Notos: Building a Dynamic Reputation System for DNS

Manos Antonakakis, Roberto Perdisci , David Dagon, Wenke Lee, and Nick Feamster

College of Computing

Georgia Institute of Technology

Atlanta, Georgia

ONR MURI

Review Meeting

June 10, 2010

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
10 JUN 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Notos: Building a Dynamic Reputation System for DNS	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Georgia Institute of Technology,College of Computing,Atlanta,GA,30332**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
MURI Review, June 2010. U.S. Government or Federal Rights License

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**17**	

Problems with Static Blacklisting

- Malware families utilize large number of domains for discovering the “up-to-date” C&C address
  - Examples are the Sinowal, Bobax and Conficker bots families that generate tens of thousands new C&C domains every day
  - IP-based (dynamic or not) blocking technologies cannot keep up with the number of IP addresses that the C&C domains typically use
  - DNSBL based technologies cannot keep up with the volume of new domain names the botnet uses every day
- Detecting and blocking such type of **agile botnets** cannot be achieve with the current state-of-the-art

Outline

- Notos
  - Notations, Passive DNS trends, and anchor-zones
  - Network based profile modeling
  - Network and zone based profiles clustering
  - Reputation function
  - System implementation
  - Results
- Conclusions and Future Work

Notos

- Network and zone based features that capture the characteristics of resource provisioning, usages, and management by domains.
  - Learn the models of legitimate and malicious domains
- Classify new domains with a very low FP% (0.3846%) and high TP% (96.8%).
  - Days or even weeks before they appear on static blacklists.

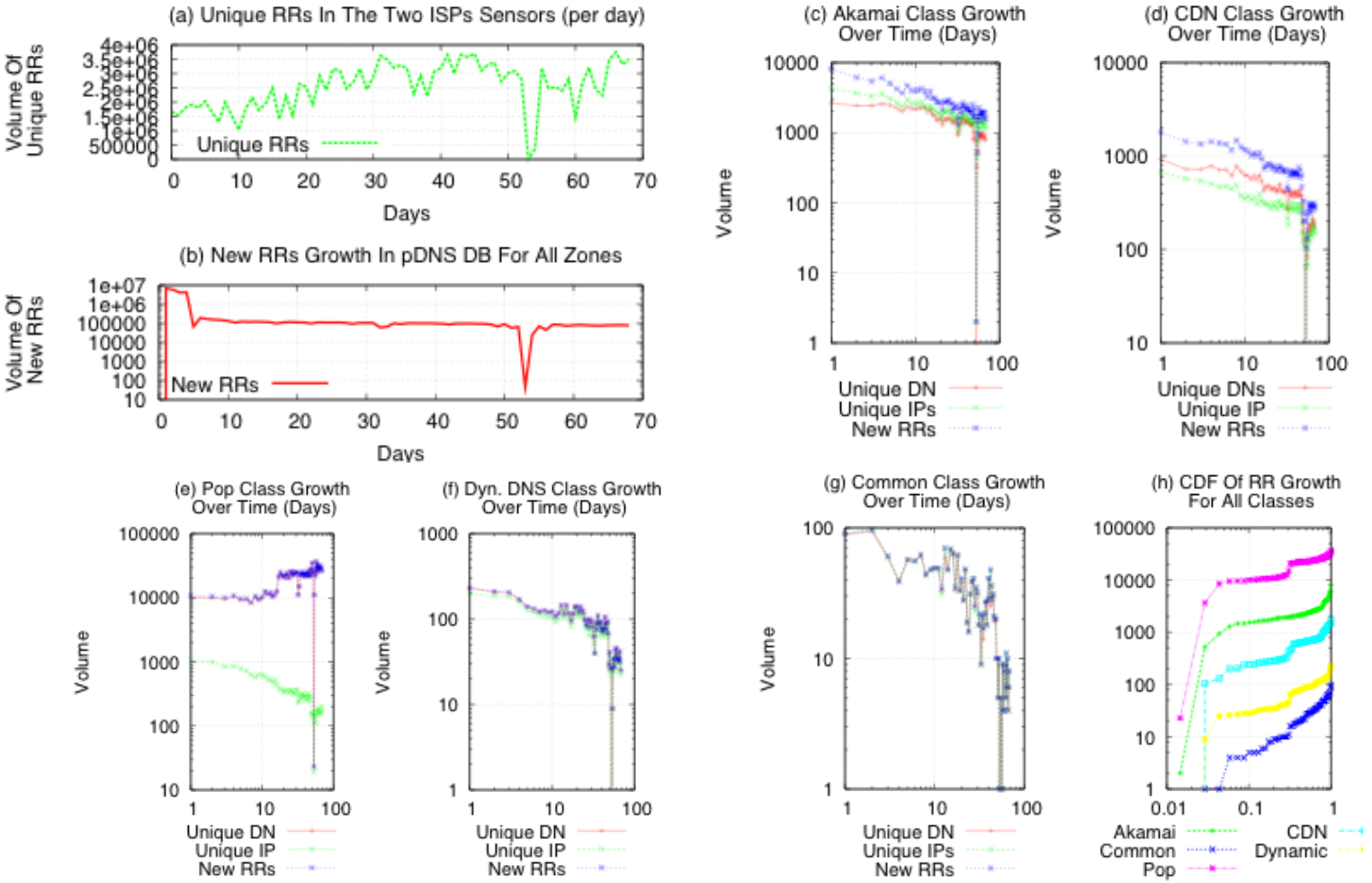
Notation & Terminology

- Resource Record (RR)
  - www.example.com 192.0.32.10
- 2nd level domain (2LD) and 3rd level domain (3LD)
  - For the domain name www.example.com: 2LD is the example.com and 3LD is the www.example.com
- Related Historic IPs (RHIPs)
  - All “routable” IPs that historically have been mapped with the domain name in the RR, or any domain name under the 2LD and 3LD
- Related Historic Domains (RHDNs)
  - All fully qualified domain names (FQDN) that historically have been linked with the IP in the RR, its corresponding CIDR and AS

Passive DNS data

- Successful DNS resolutions that can be observed in a given network
- Data set has traffic from 2 ISP sensors - one in west coast and one in east coast, also data from SIE
- We observe that different classes of zones demonstrate different passive DNS behaviors
- The number of new domain names and IPs we observe every day is in the range of 150,000 to 200,000

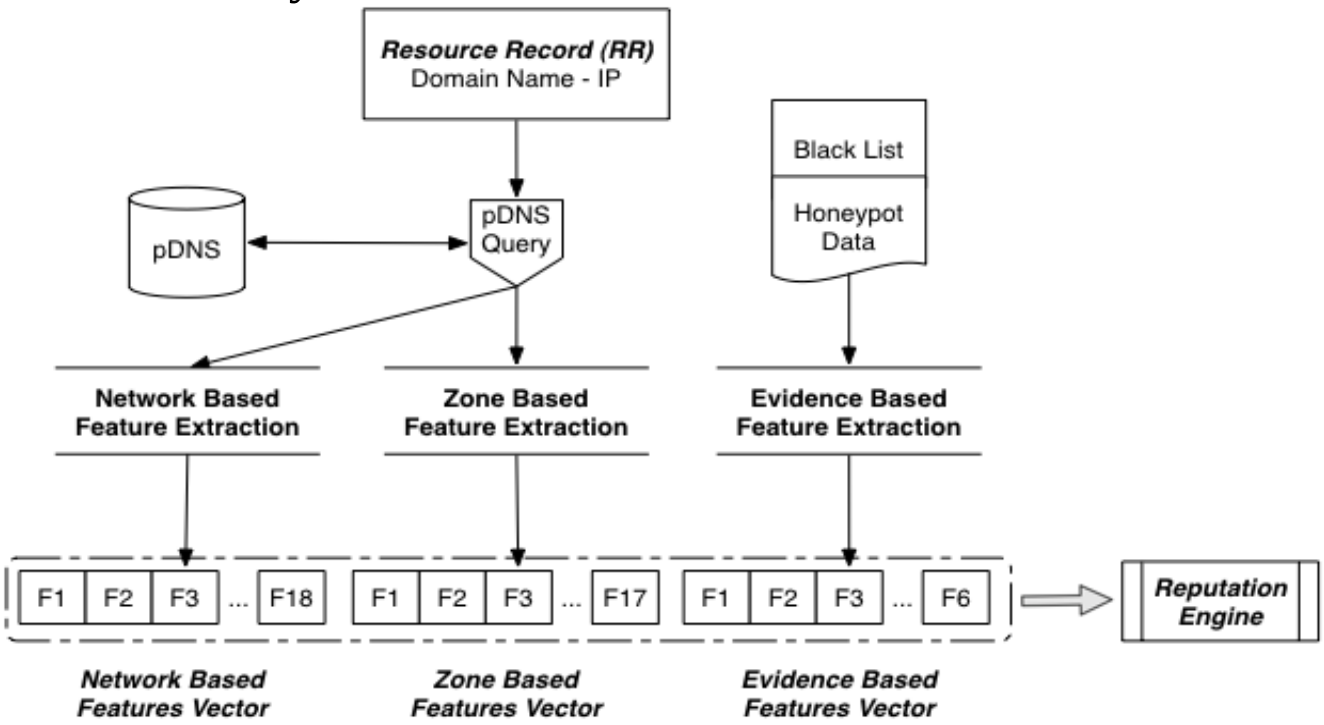
Passive DNS trends



Anchor classes in pDNS: Akamai, CDN, Popular, DYNDNS and Common

Features

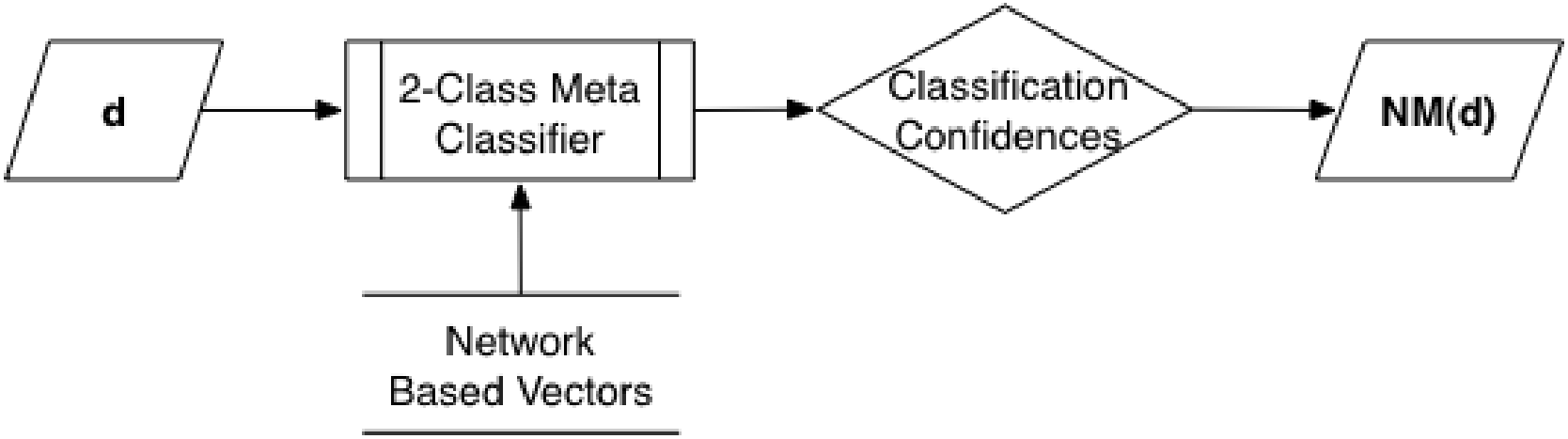
Notos computes three feature vectors for a RR, based on its RHIPs, RHDNs and Evidence data. The analysis of these feature vectors is forwarded to the reputation engine.



These 3 vectors are the Network Based Feature Vector [18], Zone Based Feature Vector [17] and the Evidence Based Feature Vector [6].

Network Profile Modeling

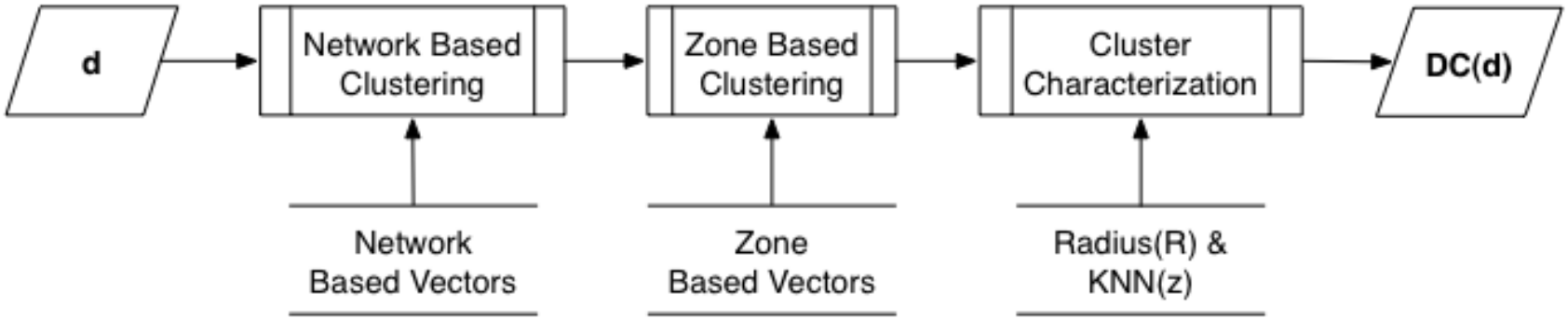
- Train a Meta-Classifier based on the 5 anchor-classes
- The network feature vector of a domain name *d* is translated into the network modeling output *(NM(d))*



The *NM(d)* is a feature vector composed from the confidence scores for each different anchor-class

Domain Clustering

The network and zone based feature vectors of a domain *d* are used to produce the domain clustering output *(DC(d))*



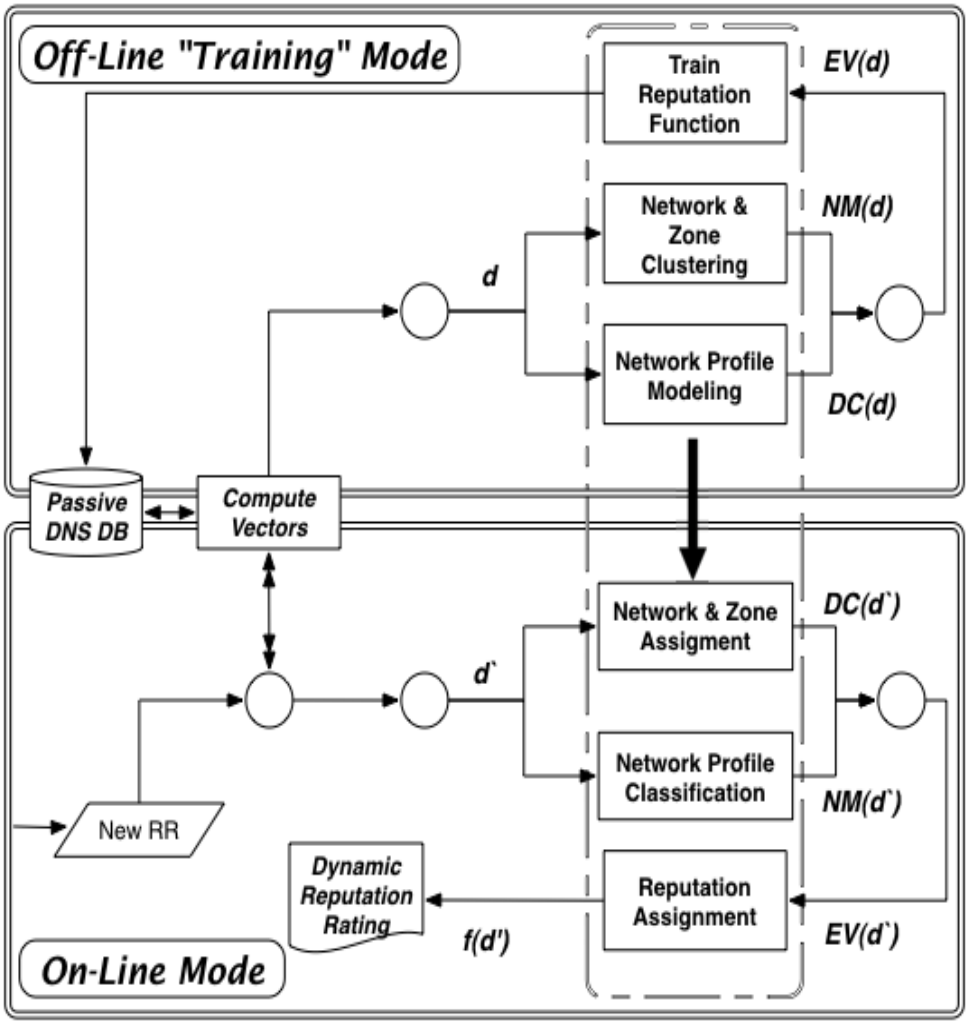
In this step we are able to **characterize** unknown domains within clusters based on already labeled domains **in close proximity**. The *DC(d)* is a 5-feature vector characterizing the position of *d* in the cluster.

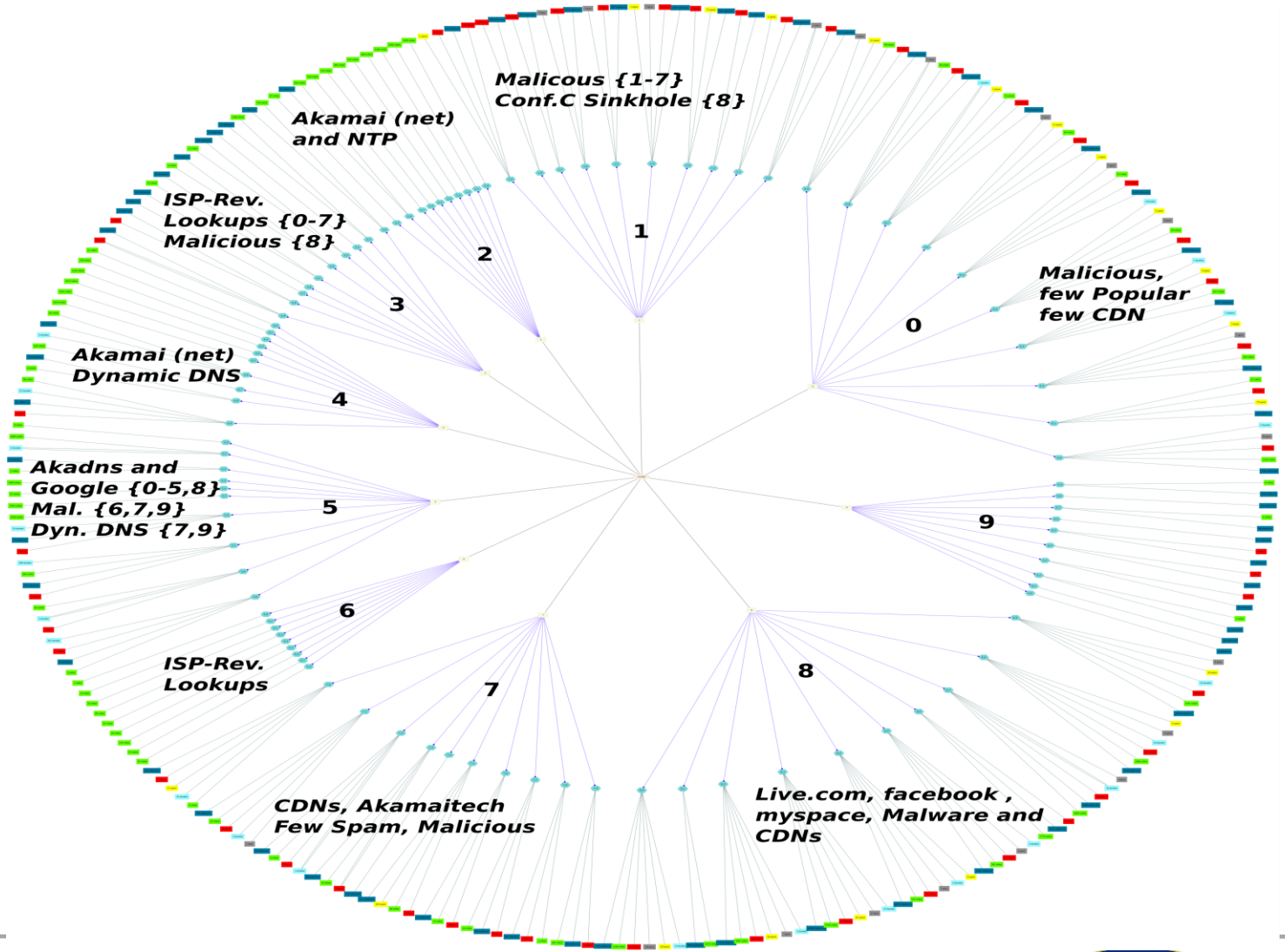
Reputation Function

- Each domain *d* in our dataset is transformed into three feature vectors by Notos: *NM(d)*, *DC(d)* and *EV(d)* (evidence profile output); these vectors assemble the reputation vector *v(d)*
- The reputation function *f(v(d))* assigns a score to the domain name *d* between [0,1]
- The reputation function is a statistical classifier (Decision Tree with Logistic Boost - after model selection)
- The reputation function is trained using labeled domain data

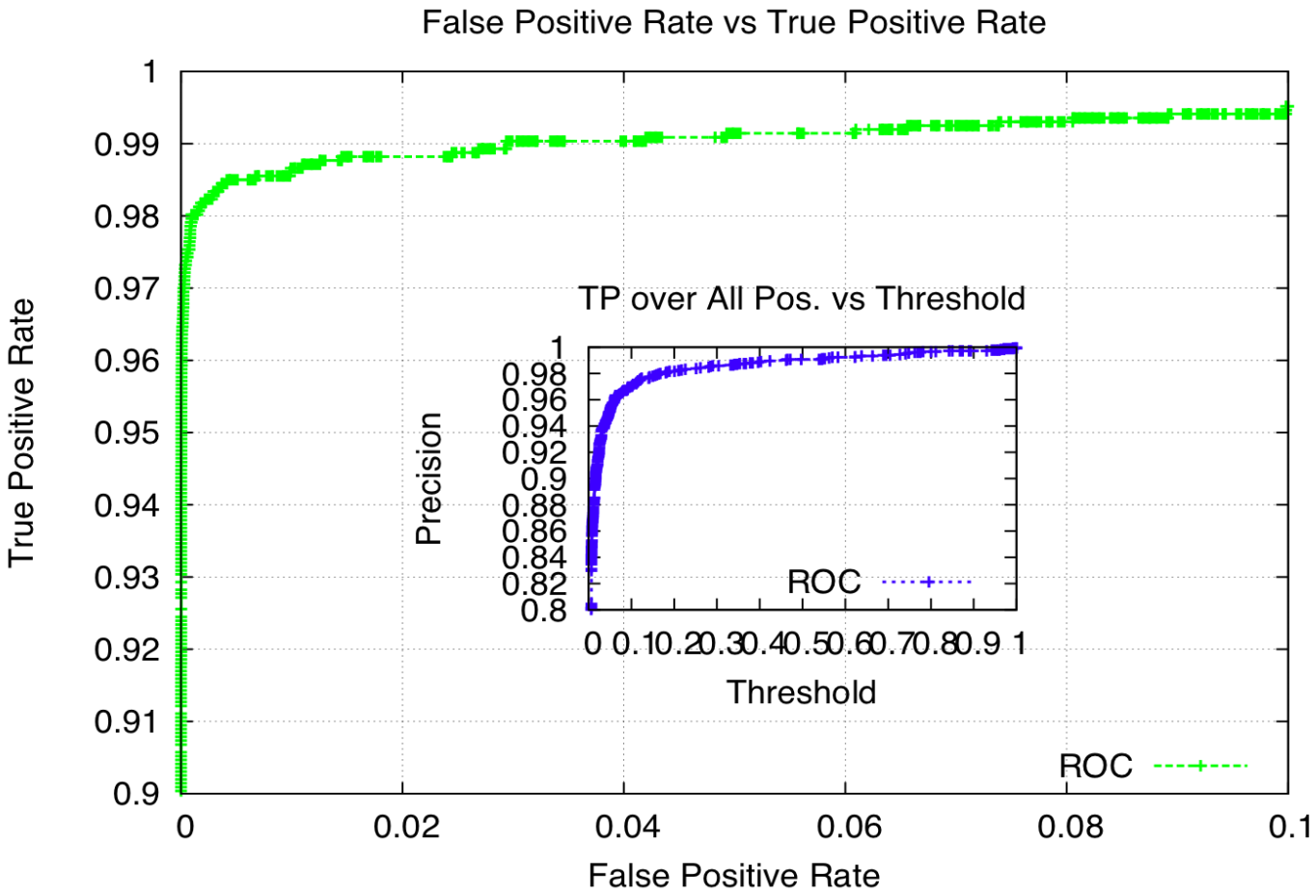
Operational Model of Notos

- Notos utilizes the **Off-line mode** to train classifiers, build the clusters and train the reputation function
- In the **In-line mode**, Notos assigns reputation to new RRs observed at the monitoring point




Malicous {1-7}
Conf.C Sinkhole {8}
Akamai (net)
and NTP
ISP-Rev.
Lookups {0-7}
Malicious {8}
Malicious,
few Popular
few CDN
Akamai (net)
Dynamic DNS
Akadns and
Google {0-5,8}
Mal. {6,7,9}
Dyn. DNS {7,9}
ISP-Rev.
Lookups
CDNs, Akamaitech
Few Spam, Malicious
Live.com, facebook ,
myspace, Malware and
CDNs
1
2
3
4
5
6
7
8
9
0

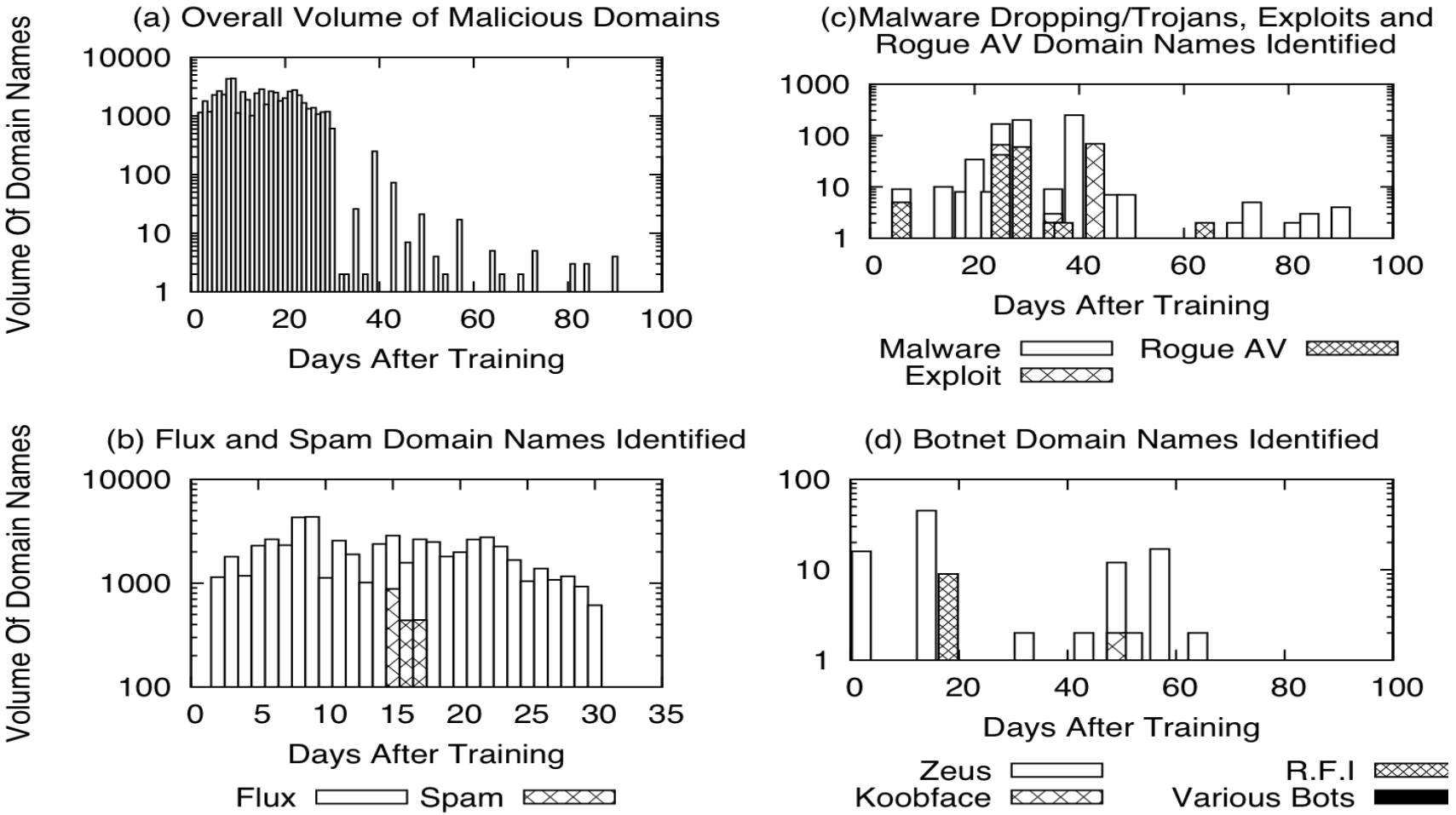
Results from the Reputation Function



FP%=0.3849% and TP%=96.8%

Results from the Reputation Function (cont'd)

of days the detection earlier than public BLs



Tech Transfer

- Damballa is actively evaluating Notos
- ISPs are interested in having us extend this line of research
- DNS vendors and other network operators
  - Have been spending millions of $ and years trying to build similar system, but fail to match Notos' capability/performance
  - Trying to get Notos technologies

Conclusions and Future Work

- Conclusions:
  - Combining network, zone, and evidence features provides the ability to dynamically associate unknown domains to known domains/networks
    - Benefits: with limited labeled domains we can identify new malicious ones, **much sooner than BLs**
- Future Work:
  - Targeted detection: use an additional clustering step based on association with specific fraudulent domain name class (RBN, Zeus, etc.) to enable targeted detection
  - Combine Notos with Spam/Flux detection systems